

# NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB

Jurgen F. Doreleijers<sup>1,2,\*</sup>, Wim F. Vranken<sup>3</sup>, Christopher Schulte<sup>4</sup>, John L. Markley<sup>4</sup>, Eldon L. Ulrich<sup>4</sup>, Gert Vriend<sup>2</sup> and Geerten W. Vuister<sup>1,5,\*</sup>

<sup>1</sup>IMM, Radboud University Nijmegen, Geert Grooteplein 26-28, 6525 GA Nijmegen, The Netherlands, <sup>2</sup>CMBI, Radboud University Nijmegen Medical Centre, Geert Grooteplein 26-28, 6525 GA Nijmegen, The Netherlands, <sup>3</sup>Department of Structural Biology, VIB and Structural Biology Brussels, Vrije Universiteit Brussel, Building E, 4th Floor, Pleinlaan 2, 1050 Brussels, Belgium, <sup>4</sup>BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Dr., Madison, WI, 53706, USA and <sup>5</sup>Department of Biochemistry, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 9HN, UK

Received October 19, 2011; Revised November 3, 2011; Accepted November 8, 2011

## ABSTRACT

For many macromolecular NMR ensembles from the Protein Data Bank (PDB) the experiment-based restraint lists are available, while other experimental data, mainly chemical shift values, are often available from the BioMagResBank. The accuracy and precision of the coordinates in these macromolecular NMR ensembles can be improved by recalculation using the available experimental data and present-day software. Such efforts, however, generally fail on half of all NMR ensembles due to the syntactic and semantic heterogeneity of the underlying data and the wide variety of formats used for their deposition. We have combined the remediated restraint information from our NMR Restraints Grid (NRG) database with available chemical shifts from the BioMagResBank and the Common Interface for NMR structure Generation (CING) structure validation reports into the weekly updated NRG-CING database (<http://nmr.cmbi.ru.nl/NRG-CING>). Eleven programs have been included in the NRG-CING production pipeline to arrive at validation reports that list for each entry the potential inconsistencies between the coordinates and the available experimental NMR data. The longitudinal validation of these data in a publicly available relational database yields a set of indicators that can be used to judge the quality of every macromolecular structure solved with NMR. The remediated NMR

experimental data sets and validation reports are freely available online.

## INTRODUCTION

Experimentally determined biomacromolecular three-dimensional (3D) structures typically are deposited in the Worldwide Protein Data Bank (wwPDB) (1–3) as a requirement by most journals including *NAR*. As of September 2011, there were over 76 000 entries in the PDB (*cf.* Table 1) of which ~9000 entries had been solved by NMR. The BioMagResBank (BMRB) (4) serves as a global repository of experimental NMR data, such as restraints, assigned chemical shifts and dynamic order parameters. Together, these repositories present a valuable resource for numerous research areas in the life sciences.

A series of experiments have shown that many NMR structures can be improved if they are recalculated from the original experimental data using present-day software and refinement protocols (5–7) including the STAP database published in this ‘Database’ issue of *Nucleic Acids Research*. These efforts have revealed that the deposited experimental data were highly heterogeneous in format, completeness and quality. Recently, we performed a large-scale optimization of X-ray derived PDB entries (8), which showed that nearly three quarters of these could be improved in terms of fit with the experimental data and geometric quality (9). The massive scale of this effort also allowed the analysis of even the smallest improvements in a statistically meaningful way (10).

\*To whom correspondence should be addressed. Tel: +44 116 229 7076; Fax: +44 116 229 7018; Email: [gv29@le.ac.uk](mailto:gv29@le.ac.uk)  
Correspondence may also be addressed to Jurgen F. Doreleijers. Tel: +31 24 36 19674; Fax: +31 24 36 19395; Email: [jurgenfd@gmail.com](mailto:jurgenfd@gmail.com)

**Table 1.** PDB entries

Set	Entries
PDB	76 003
Solution NMR	9042
NRG-CING	8915
Proteins	7967
Dimers	413
Complexes	1235
Ligands	384
Deposition	
Before 1990	9
1990-2000	1920
After 2000	7113

Overview of subsets of PDB entries (23 September 2011).

Recalculation and proper validation (i.e. validation including the experimental data) both require that the underlying experimental data are syntactically and semantically correct. We have therefore worked for several years on this topic (11,12). In collaboration with the BMRB, we have completed the remediation of the NMR restraint data entries, which resulted in the NMR Restraints Grid (NRG) databases. We recently added the BMRB chemical shift (CS) data and these combined results have been subjected to our integrated NMR structure and experimental data validation analyses, to yield the new database described in this contribution. We have named this database NRG-CING. The database is freely available at <http://nmr.cmbi.ru.nl/NRG-CING> and it will be updated on a weekly basis. For the NRG-CING pipeline, we have extended the Common Interface for NMR structure Generation (CING; pronounced 'king') software package (G. Vuister, *et al.*, CING; an integrated residue-based structure validation program suite, manuscript in preparation). The pipeline first assembles a set of experimental and structural data and then produces a report that includes the results of eleven computer programs that were written by us or by others. The quality of the structure coordinates is currently determined mainly by WHAT\_CHECK (12) and PROCHECK-NMR (13). The experimental restraints are tested for consistency and agreement with the structure by CING, Wattos (14), and PROCHECK-NMR/Aqua (13). In addition, the systematic analysis of NMR restraints allowed us to extract new patterns of recurring problems (15). Validation of CS values based on structural and sequence information by CING and the external programs VASCO (16) and SHIFTX (17) and TALOS+ (18) is an integral part of the analyses.

The NRG-CING database is a coherent, annotated and verified collection of experimental input data, the resulting structures and the analyses of their quality. NRG-CING will be the basis for recalculation efforts such as the STAP (<http://psb.kobic.re.kr/stap/refinement>) and LOGRECOORD (7) databases that will lead to better quality NMR structure ensembles that in turn will allow researchers in the life sciences, in drug design and in bioinformatics to better perform their structure-based research.

## DATA PREPARATION

### Data conversion

The creation of a coherent and validated database of both structures and experimental data requires several steps. For the NRG-CING production pipeline we employed four stages, that we call C, R, S and F denoting coordinate, restraint, chemical shift and filtering, respectively (Figure 1).

*Coordinate stage.* The coordinate data flow in from the wwPDB using an mmCIF formatted file that adheres to the PDB eXchange dictionary (pdbx).

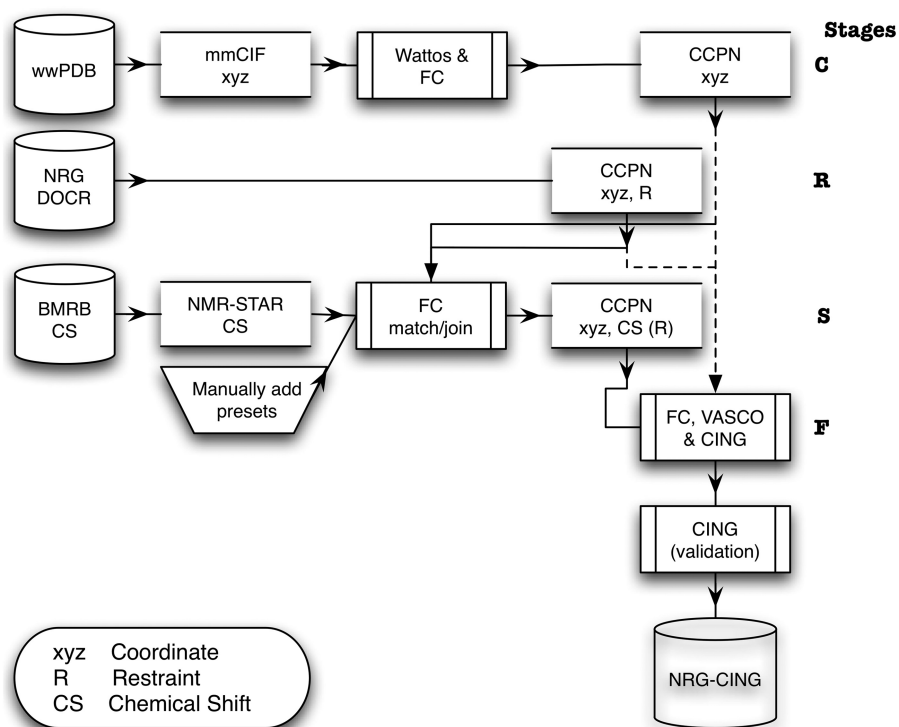
*Restraints stage.* When restraints are present, the coordinates and the restraints are imported directly from the NRG Database Of Converted Restraints [DOCR; (11)] at BMRB as a CCPN XML file.

*Shift stage.* We developed code in collaboration with BMRB to run through a wide variety of data sources in order to match older entries for which the match relation between BMRB and PDB entries had not yet been archived. The matching algorithms are documented for the NRG part at: <http://tinyurl.com/68dd9I9> and the CING part at <http://tinyurl.com/67vfuyI>. The CS data from BMRB are then merged by the FormatConverter (FC) (19) in a procedure similar to the one used for the restraints (15).

*Filter stage.* The distance restraints (DR) are stereospecifically checked and in some cases corrected by FC and CING using the same method as currently in use at the BMRB (11). Distance restraints with violations over 2 Å (up to a maximum of three per entry) were omitted from the NRG-CING database and are labelled as outliers. Although such DRs are sometimes correct, the impact of removing correct DRs is deemed to be less detrimental compared to the effects of retaining potentially incorrect ones. In particular, the latter situation could result in unjustified labelling of an entry to be in discord with its experimental data. From anecdotal interactions with depositors we know that these restraints are often errant violations that were not observed at the time of structure calculation, but arose later as a consequence of correcting other problems, for example, typographical errors that led to a restraint being accidentally uncommented or incorrect mapping of one or two atom names. The referencing of the CS is validated during this stage by VASCO, which compares the CS values for the atoms in a protein to their statistical distribution in relation to the coordinate-derived per-atom solvent exposure (16).

### Cloud computing

The CING calculations require on average 20 min per entry for a total of 3000 core hours to process the current set of entries. Most of that time is used to run the many external programs and to prepare the large number of plots that report on the data. Because the complete database needs to be reassembled following each major overhaul of the analysis, this project continues



**Figure 1.** Flow chart. Data flow chart showing the software tools involved in this project: CING, Wattos and FormatConverter (FC). The four stages denoted: C, R, S and F are described in the text. The dashed line indicates an alternative to the default route including all data types. The repositories, programs and data-formats are represented by cylinders, 'closed rectangles' and 'open rectangles', respectively.

to require substantial computing power. As CING has many external program dependencies, it cannot easily be installed on a traditional grid, but we have found it to be very suitable for a cloud computing setup. The eleven programs required for generating a CING report besides CING (G. Vuister *et al.*, manuscript in preparation) are: CCPN (19), DSSP (20), Matplotlib (<http://matplotlib.sourceforge.net>), MOLMOL (21), PROCHECK/Aqua (13), Povray (<http://www.povray.org>) ShiftX (22), TALOS+ (18), VASCO (16), Wattos (14) and WHAT\_CHECK (12). We use the cloud facilities at SARA, our industrial partner Bitbrains (Amstelveen, NL, USA) and WeNMR/INFN for each full iteration in the NRG-CING project.

### Project management

A large international collaborative project like NRG-CING requires the identification and remediation of issues with software developed and procedures used. From the beginning of this project in 2008, the issues were maintained in a Google Code repository at <http://code.google.com/p/cing> and linked to the source code in the CING project. Together with the general CING issues, almost all of the 300+ issues currently listed have been addressed. The documentation is described in Wiki pages at the same site. An automatic build and test farm for several Operation Systems is managed by Jenkins Continuous Integration (CI, <http://jenkins-ci.org>) at <http://nmr.cmbi.ru.nl/jenkins/job/CING>.

## RESULTS

### NRG-CING database overall composition

Of the 8915 entries contained in the NRG-CING database (September 2011) 5423 contained experimental data including DRs (Tables 1 and 2). These entries span the full time frame during which NMR structures have been deposited (1988 to present). Analysis of the experimental data variation also showed that the set contains structures determined both from 'sparse data', where only a limited amount of structural information was extracted from NMR experiments, and from abundant experimental data.

### Examples of longitudinal validation

The CS values of the  $\beta$  and  $\gamma$  carbons of proline have been shown sensitive to the usual *trans* or the occasionally occurring *cis* peptide bond configurations. A study based on 33 *cis* and 1000 *trans* Pro residues in non-paramagnetic proteins showed a clear clustering for the  $^{13}\text{C}$   $\beta/\gamma$  CS difference (CSD) values (23). The regions of (0.0, 4.8) and (9.15, 14.4) ppm corresponded with near absolute certainty to the *trans* and *cis* conformations, respectively. In NRG-CING we observe 228 *cis* and 7949 *trans* Pro in 3435 entries with  $\beta/\gamma$  carbons CS values obtained from BMRB. We have identified the reversed correspondence for 8 (*cis*) and over 100 (*trans*) occurrences. For example, the recent Structural Genomics PDB entry 2k8s (Cort J.R. *et al.*, unpublished results) Pro57 in

**Table 2.** Statistics of the NRG-CING database

Set	Entries	Per entry count		
		Average (SD)	Min.	Max.
Experimental restraints	5519	1392 (1158)	9	11 044
Distances (DRs)	5423	1325 (1107)	11	10 112
AIR DRs only <sup>a</sup>	97	27 (14)	11	49
Dihedral angles <sup>b</sup>	3401	128 (106)	9	1099
RDCs	426	139 (148)	9	970
Chemical shifts	3626	780 (512)	2	3959
Number of residues	NA	92 (71)	2	1659

<sup>a</sup>The number of HADDOCK AIR entries was overestimated by including every NRG-CING entry with <50 DRs. <sup>b</sup>The number of entries with dihedral angle restraints is overestimated by including CS derived ones from Talos+.

NA: Not Applicable.

chain B has a CSD of 11.9 ppm, which indicates a contradiction with the *trans* state modelled in all conformers of the ensemble. We also observed much more extreme CSD values that are likely caused by human error: e.g. the CSD of Pro71 in PDB entry 2i4k (24) has a very large value (37 ppm) that most likely resulted from uncorrected folding/aliasing of the NMR spectrum.

A second example of the combined analysis of chemical shifts in relation to structural quality concerns the sidechain conformation of the leucine delta carbons. Also here, chemical shifts have proven reliable indicators of conformation (25). For the NRG-CING database, 218 (*trans*) and 115 (*gauche*<sup>+</sup>) structured leucine residues in a total of 286 entries showed inconsistencies between observed chemical shifts and  $\chi_1/\chi_2$  sidechain conformations, that warrant further investigation (Berntsen, K.R.M. Doreleijers, J.F., Breukels, V., Stens, E., Vriend, G. and Vuister, G.W. manuscript in preparation).

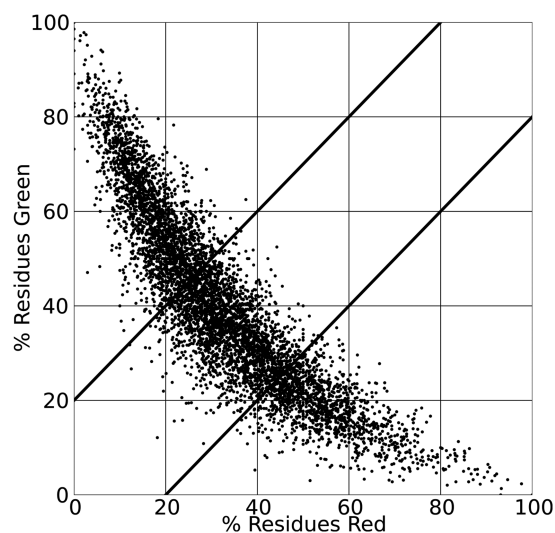
## AVAILABILITY

### Reports

Currently all wwPDB members (RCSB-PDB, PDBe, PDBj and BMRB) include links to the NRG-CING reports. These pointers drive the vast majority of traffic to the NRG-CING database. The complete NRG-CING database can be accessed by any user. In addition to straightforward selection of specific PDB entries, the front page of the NRG-CING website also allows interactive selection using different criteria, such as protein size, number of distance restraints or chemical shift restraints or ROG score. According to Google Analytics, during the last year NRG-CING was on average visited each day ~25 times by 9 'absolute unique visitors'.

### Relational database

In addition to the web-based interactive HTML, CSV dumps from the relational database are available (<http://nmr.cmbi.ru.nl/NRG-CING/pgsqli>). These files can be imported to a slave database using the SQL script at <http://tinyurl.com/3rb24eq>. The relational database (RDB) contains the validation data at the levels of entry, chain, residue and atom with special tables



**Figure 2.** ROG Results from NRG-CING. The percentage of residues with ROG score red (bad) versus green (good) is plotted with filled circles for 6265 NMR PDB entries from NRG. The red, orange, green (ROG) score is a composite assessment over individual program's validation criteria on the quality of entities such as restraint, coordinate, peak, chemical shift, atom, residue, molecule, etc. The ROG scores are propagated based upon defined relationships between such entities. The entries were selected to have at least: 3.5 kDa molecular mass, 10 models and one protein chain. On the bottom right of the banana-shaped distribution are a minority of entries that have a significant fraction of residues marked red. Note that the percentages green and red taken together with the omitted dimension for orange, add up to 100%.

recently added for DRs and CSs. Many of the validation criteria in CING are also in this relational database, and plots are available at <http://nmr.cmbi.ru.nl/NRG-CING/HTML/plot.html>, showing the distribution of values such as detailed in Figure 2 for the CING ROG scores. The NRG-CING RDB is setup in conjunction with the PDBj Mine RDB for full cross-correlated access to PDB meta data such as deposition dates (26).

### iCing Server and service

Our multilingual web server (<https://nmr.cmbi.ru.nl/icing/>) and a web service together are called iCing (see Figure 3). It allows a user to submit NMR-derived coordinates, restraints and CS values in three data formats. We preferentially employ CCPN project files (19), but also accommodate additional data formats, such as the out-dated plain PDB format for structural data only. Although not preferred, this capability does provide the casual user access without sophistication. In collaboration with Dr Torsten Herrmann, we added the capability to upload CYANA formatted data, which will facilitate more standalone programs to integrate with the iCing service.

The iCing server can be used prior to a submission or, even better, as part of the iterative process of NMR structure determination. Figure 3 shows that the user can customize the validation criteria, which can be useful to specifically focus attention on particular aspects.





**Figure 3.** iCing Web Server and Service. The screenshot of iCing (Spanish translation selected) shows the customizable definitions for 'poor' (orange) and 'bad' (red) that CING will use for some WHAT\_CHECK parameters. The Google Web Toolkit (GWT) allowed us to easily add German, Spanish, French, Italian, Japanese, Dutch, Portuguese, Russian and Chinese translations to the default English language with help from our colleagues who are native speakers of these languages.

Generally speaking, however, this is not recommended because the standard criteria are used in deriving the NRG-CING database. Validation of the validation criteria themselves is a topic of ongoing research.

The server uses a simple three-tier setup with a Google Web Toolkit 2.0 front end, an Apache/Tomcat secured HTTP servlet, and a backend part including the CING installation. The iCing server has seen 1025 unique views during the first 10 months of 2011, according to Google Analytics. The standalone CCPN Analysis program (19) is using iCing as a service extensively. In total, the iCing service has been used for 1417 data sets in the same period.

## FUTURE PERSPECTIVES

### Improvements

Although already a valuable resource, as judged from its usage statistics, we continuously seek improvements to the database. We plan to address the following topics: (i) we aim to make the database 100% complete by solving a series of difficult data-related issues (such as Google Code NRG issue 272 and CING issues 266, 310–312) that currently limit us to include only 98.6% of the PDB entries. (ii) We plan on improving the NRG-CING setup with better matches between older BMRB and PDB entries, deposited before the relationship between these was maintained. (iii) Finally, although RDC data are contained within the database, these should be validated as well.

### Usage

Finally, NRG-CING only contains the released PDB entries. This journal, *Nucleic Acids Research* like many

journals, encourages authors of new structure papers to provide referees with the output from PDB's validation report from <http://deposit.pdb.org/validate>. It would be of great value to authors and referees to have these CING reports available in addition to the currently used validation reports on the coordinates alone.

## ACKNOWLEDGEMENTS

We gratefully like to acknowledge the iCing translators: Jundong Lin (Madison), Alan W.S. da Silva (Cambridge), Thomas Lütke (Gießen, Germany) Philip Kensch (Nijmegen), Oli Maria Victoria Grober (Naples), Naohiro Kobayashi (Osaka) and Nadia Kovalevskaya (Nijmegen). We enjoyed the advice and constructive criticism from all CCPN developers in Cambridge and Tim te Beek, Karen Berntsen, Vincent Breukels, Maarten Hekkelman, Wilmar Teunissen and Wouter Touw in Nijmegen. We appreciated Akira Kinjo's advice for setting up a local slave of PDBj Mine. We are especially indebted to all the authors providing wwPDB with the fruits of their labor. We appreciated the support from Floris Sluiter and colleagues at Sara (Amsterdam), Gjalt van Rutten at Bitbrains and Marco Verlatto (WeNMR) with using their HPC Cloud infrastructure.

## FUNDING

The Netherlands Organization for Scientific Research (NWO) (grant 700.55.443, to G.W.V.), Netherlands Bioinformatics Centre (NBIC); EU FP6 grants STREP Extend-NMR (LSHG-CT-2005-018988) and EMBRACE (LHSG-CT-2004-512092); FP7 WeNMR (grant 261572, to J.F.D., G.V. and G.W.V.); Brussels Institute for Research and Innovation (Innoviris) (grant BB2B 2010-1-12, to W.V.F.); US National Library of Medicine (grant LM05799, to C.S., E.L.U. and J.L.M.). Funding for open access charge: Radboud University Medical Centre Funds to the CMBI.

*Conflict of interest statement.* None declared.

## REFERENCES

- Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
- Velankar, S., Alhroub, Y., Alili, A., Best, C., Boutselakis, H.C., Caboche, S., Conroy, M.J., Dana, J.M., van Ginkel, G., Golovin, A. *et al.* (2011) PDBE: Protein Data Bank in Europe. *Nucleic Acids Res.*, **39**, D402–D410.
- Berman, H., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
- Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
- Nabuurs, S.B., Nederveen, A.J., Vranken, W., Doreleijers, J.F., Bonvin, A.M.J.J., Vuister, G.W., Vriend, G. and Spronk, C.A.E.M. (2004) DRESS: a database of Refined solution NMR structures. *Proteins*, **55**, 483–486.
- Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A.E.M., Nabuurs, S.B., Güntert, P., Livny, M.,

- Markley, J.L., Nilges, M. *et al.* (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*, **59**, 662–672.
7. Bernard, A., Vranken, W.F., Bardiaux, B., Nilges, M. and Malliavin, T.E. (2011) Bayesian estimation of NMR restraint potential and weight: a validation on a representative set of protein structures. *Proteins*, **79**, 1525–1537.
8. Joosten, R.P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A.-C., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Costa, A.L.D., Deleage, G. *et al.* (2009) PDB\_REDO: automated re-refinement of X-ray structure models in the PDB. *J. Applied Crystallogr.*, **42**, 376–384.
9. Joosten, R., Womack, T., Vriend, G. and Bricogne, G. (2009) Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. *Acta Cryst. D*, **65**, 176–185.
10. Touw, W.G. and Vriend, G. (2010) On the complexity of Engh and Huber refinement restraints: the angle  $\tau$  as example. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 1341–1350.
11. Doreleijers, J.F., Vranken, W.F., Schulte, C., Lin, J., Wedell, J.R., Penkett, C.J., Vuister, G.W., Vriend, G., Markley, J.L. and Ulrich, E.L. (2009) The NMR restraints grid at BMRB for 5,266 protein and nucleic acid PDB entries. *J. Biomol. NMR*, **45**, 389–396.
12. Hooft, R., Vriend, G., Sander, C. and Abola, E. (1996) Errors in protein structures. *Nature*, **381**, 272.
13. Laskowski, R., Rullmann, J. and MacArthur, M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
14. Doreleijers, J.F., Nederveen, A.J., Vranken, W. and Lin, J. (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J. Biomol. NMR*, **32**, 1–12.
15. Vranken, W. (2007) A global analysis of NMR distance constraints from the PDB. *J. Biomol. NMR*, **39**, 303–314.
16. Rieping, W. and Vranken, W.F. (2010) Validation of archived chemical shifts through atomic coordinates. *Proteins*, **78**, 2482–2489.
17. Zhang, H., Neal, S. and Wishart, D. (2003) RefDB: a database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, **25**, 173–195.
18. Shen, Y., Delaglio, F., Cornilescu, G. and Bax, A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, **44**, 213–223.
19. Vranken, W., Boucher, W., Stevens, T., Fogh, R., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, **59**, 687–696.
20. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
21. Koradi, R., Billeter, M. and Wüthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 51–55.
22. Neal, S., Nip, A., Zhang, H. and Wishart, D. (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, **26**, 215–240.
23. Schubert, M., Labudde, D., Oschkinat, H. and Schmieder, P. (2002) A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on 13C chemical shift statistics. *J. Biomol. NMR*, **24**, 149–154.
24. Zhong, Q., Watson, M.J., Lazar, C.S., Hounslow, A.M., Waltho, J.P. and Gill, G.N. (2005) Determinants of the endosomal localization of sorting nexin 1. *Mol. Biol. Cell*, **16**, 2049–2057.
25. Mulder, F. (2009) Leucine side-chain conformation and dynamics in proteins from 13C NMR chemical shifts. *Chem. Bio. Chem.*, **10**, 1477–1479.
26. Kinjo, A.R., Yamashita, R. and Nakamura, H. (2010) PDBj Mine: design and implementation of relational database interface for Protein Data Bank Japan. *Database*, **2010**, doi:10.1093/database/baq021.